DOCUMENT RESUME

ED 398 247 TM 025 187

AUTHOR Chang, Lei
TITLE A Comparison between the Nedelsky and Angoff Standard-Setting Methods.
PUB DATE Apr 96
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Cutting Scores; Difficulty Level; Distractors (Tests); *Graduate Students; Graduate Study; Higher Education; *Interrater Reliability; Knowledge Level; *Research Methodology; *Standards; Test Items
IDENTIFIERS *Angoff Methods; *Nedelsky Method; Standard Setting

ABSTRACT

It was hypothesized that, when compared to the Angoff method (W. H. Angoff, 1971), the Nedelsky method (L. Nedelsky, 1954) for standard setting had lower intrajudge inconsistency, lower cutscores, and lower cutscores especially for items presenting challenges to the judges. These hypotheses were tested and supported in a sample of 22 graduate students serving as judges to determine average performance level for 9 research methodology items using both methods. The more consistent Nedelsky methods are attributed to judges' focused use of response options as a consistent source of information. A close scrutiny of the plausibility and similarity of test item distractors tests the Nedelsky judges' own item knowledge, which, when challenged, results in a low item difficulty estimate. (Contains 3 tables and 27 references.) (Author/SLD)

A Comparison Between the Nedelsky and Angoff Standard-Setting Methods

Lei Chang

University of Central Florida

Send inquiries to Lei Chang, College of Education, University of Central Florida, Orlando, FL 32816-1250.

Paper presented at the Annual meeting of the National Council on Measurement in Education, New York City, April, 1996

Abstract

It was hypothesized that the Nedelsky versus the Angoff methods (1) had lower intrajudge inconsistency, (2) lower cutscores, and (3) lower cutscores especially for items presenting a challenge to the judges. These hypotheses were tested and supported in a sample of 22 graduate students serving as judges to determine average performance level (not MPL) for nine research methodology items using both methods. The more consistent Nedelsky decisions are attributed to judges' focused use of response options as a consistent source of information. A close scrutiny of the plausibility and similarity of distractors subjects to a test the Nedelsky judges' own item knowledge which, when challenged, results in a low item difficulty estimate.

A Comparison Between the Nedelsky and Angoff Standard-Setting Methods

A large number of studies have been conducted to compare various standard-setting procedures, particularly, the Nedelsky (1954) and Angoff (1971) methods. The results are mixed. Table 1 summarizes the results of nine Nedelsky-Angoff comparison studies identified in the past 30 years of psychometric literature. As Kane (1994) pointed out, except for alerting the public to the large discrepancies in passing scores associated with different methods, the research has been inconclusive as to which method should be preferred. The present study intended to draw a more conclusive comparison between the Nedelsky and Angoff methods by making two improvements.

Improvements of the Present Study

First, the inconclusiveness of existing findings is due to the lack of internal (Plake, Melican, & Mills, 1991) or external (Kane, 1994) criteria without which it is impossible to decide which method arrives at a more consistent and adequate passing score (Kane, 1994). The best way to check for internal consistency or intrajudge consistency of a standard-setting process is to compare the judge rated minimum performance level (MPL) with empirical item difficulty based on minimally competent examinees. However, p-values of minimally competent examinees are rarely available. Compromises are suggested by using p-value estimates of the entire population (e.g., Cross, Impara, Frary, & Jaeger, 1984) or estimates for a subgroup whose performance is

considered to best approximate minimum competency (e.g., Plake et al., 1991). Both approaches have problems (see Plake et al., 1991). The present study had judges determine average performance level (APL) instead of MPL so that p-value estimates for the entire population could be used as an adequate criterion to evaluate the APLs yielded by the two contrasting methods. Although setting an APL did not exactly represent the usual standard-setting objective, this approach did not change the procedural features of the two standard-setting methods which were being compared. Thus, the external validity of the comparison was nearly not compromised. More importantly, the internal validity of the study was improved by ensuring the availability of an adequate objective criterion against which intrajudge consistency of the comparing methods could be conclusively evaluated.

Second, the present study improved the conclusiveness of a comparison by having the same judges use both the Nedelsky and Angoff methods. When different judges were employed to use different standard-setting methods, differences among methods could have been confounded by differences of judges. The possibility of such a confound is underscored by many empirical studies which have demonstrated the influence of judge characteristics in setting standards (Behuniak, 1982; Busch & Jaeger, 1986, 1990; Cross, Frary, Kelly, Small, & Impara, 1985; Jaeger, 1982; Jaeger, Cole, Irwin, & Pratto, 1980, cited in Jaeger, 1989; Roth, 1987; Plake, Impara, & Potenza, 1994;

Saunders, Ryan, & Huynh, 1981). Among these investigations, Behuniak's (1982) finding particularly warns of the danger of confounding method differences by judge differences. Behuniak (1982) had two groups of judges of three each use the Angoff procedure and another two groups of judges of four each use the Nedelsky procedure to set standards for a 9th-grade reading test. Standards from each pair of the two groups using the same method were significantly different. In fact, for some groups, the two standards obtained using the same method showed larger differences than standards obtained using different methods.

In light of Behuniak's (1982) finding, same judges should be used to draw any meaningful comparisons among standard-setting methods. As Livingston (1982) stated, "These studies must use the same judges to make both kinds of judgments, or else they will not be able to separate the effects of different judges from the effects of different methods" (p.6). Unlike some experimental treatments which, when tested in a repeated measures design, may produce unwanted carry-over effects, different standard-setting methods are not expected to influence each other. Even if the experience with one standard-setting method affects the use of another method, such influence is likely to make the two resulting standards more similar than different. The rationale behind a comparison among standard-setting procedures is, by and large, an expectation of differences among them. Thus, using a repeated measures design to investigate differences among standard-setting procedures would help protect

Type I error in addition to strengthening the validity of a comparative conclusion.

In addition to these improvements, the present study tried to develop hypotheses to postulate expected differences between the Nedelsky and Angoff methods prior to data collection.

Rationales and Hypotheses

Extending the research on decision making, Smith and Smith (1988) postulated that the Angoff and Nedelsky methods directed judges to different information when arriving at standards. They hypothesized that Nedelsky judges made use of response options almost exclusively as salient information in making decisions whereas the Angoff judges used various other sources of information. Their hypothesis was supported when tested in a sample of 31 judges who were randomly assigned to either the Angoff or Nedelsky method to set standards for a high school reading comprehension test. Their finding underscores the procedural difference between the two methods -- the Nedelsky judges are instructed to evaluate the response alternatives whereas the Angoff method does not restrict the judges to specific features of an item. Because differences in the framing of a task result in differential use of available information (Smith & Smith, 1988), response option information, as shown in Smith and Smith's study, is not used exclusively in reaching an Angoff decision. Although Smith and Smith (1988) implied in their discussion that the Angoff method may prove more adequate than the Nedelsky method because of its use of more varied

information, such a conclusion, as the authors strongly cautioned, is limited to the nature of tests and items. In a reading comprehension test, for example, the length, readability, and grammatical structure of a passage may provide salient information on their own for a decision on item difficulty or minimum performance level. However, in other short-answer tests, such as science and mathematics, where a correct concept depends on the matching of the stem of an item with one correct response option, similarity and plausibility of response options represent primary factors contributing to the difficulty or easiness of the item. The Nedelsky judges' reliance on the same salient source of information in reaching decisions, as was found by Smith and Smith (1988), is conducive to the internal consistency of the Nedelsky judgmental process and will result in lower intrajudge inconsistency. On the other hand, judgmental inconsistency may possibly result from Angoff judges' use of different sources of information. Thus, for the short-answer items where the stem and a response together form a complete concept or fact, the Nedelsky decisions which are driven almost exclusively by evaluating the plausibility and distinctiveness of response options are expected to have lower intrajudge inconsistency than the Angoff ratings.

For both the Nedelsky and Angoff methods (as well as other test-centered approaches) judges have to anticipate what the targeting examinees can and can not do, an imaginary process that is much subject to error, or intrajudge inconsistency. In the Angoff method, one decision is made for each item, leaving no

room to balance out the error associated with the decision. In the Nedelsky method, on the other hand, the final rating of the item is, practically, an average of as many such error-prone decisions as there are distractors of an item. Averaging serves to reduce errors. For example, assume the truth is that the targeting examinees do not stand a chance of correctly answering a four-choice item. An Angoff judge who erroneously rated 1.0 for the item would commit an intrajudge error of 1.0. To commit the same magnitude of error, a Nedelsky judge has to make three wrong decisions by erroneously crossing out all three distractors none of which, as the truth dictates, the targeting examinees are able to eliminate. If the Nedelsky judge is right one of the three times, the resulting intrajudge error shall be smaller than that which is associated with the Angoff decision. Thus, by using consistent information and by making multiple decisions, the Nedelsky method was hypothesized to have lower intrajudge inconsistency than the Angoff method (Hypothesis 1).

The impact of judges' subject matter knowledge on standard-setting has been shown in several studies (Busch & Jaeger, 1990; Chang, Dzuiban, Hynes, & Olson, 1994; Cross et al., 1985; Jaeger, 1982; Pavia & Vu, 1979; Saunders et al., 1981). For example, Chang et al. (1994) found that judges tended to set high standards for items they answered correctly and low standards for items they answered incorrectly. Pavia and Vu (1979) observed that Nedelsky judges had difficulties in separating their own difficulty with items from rendering judgments on these items.

As the Nedelsky judges evaluate each response alternative of an item, their knowledge underlying the item will be probed more than that of the Angoff judges who are not required to closely review the response options. Possible difficulties Nedelsky judges may experience with an item will thus be factored into the MPL decision, resulting in a lower standard. In contrast, potential doubt Angoff judges may have about an item could be attenuated by the presence of a correct answer and the absence of a scrutiny of the nuances among the alternatives. Thus, the lack of expertise underlying an item will have less impact on an Angoff than a Nedelsky decision. To put it differently, an Angoff decision is primarily driven by confirming one correct alternative whereas a Nedelsky decision is based on disproving three (for a four-choice item) false alternatives. The judges' underlying knowledge has a higher chance of being probed in the Nedelsky than the Angoff procedure. If item-related knowledge indeed influences judges' ratings as has been shown in the literature, the Nedelsky method shall produce lower standards than the Angoff method, especially for the items with which judges have difficulties. It was hypothesized that a Nedelsky cutscore was lower than an Angoff cutscore (Hypothesis 2) and the difference was larger for items judges answered incorrectly than for items judges answered correctly (Hypothesis 3).

Method

Subjects

Subjects were 22 graduate students in education enrolled in a research method course. They learned about and practiced the Angoff and Nedelsky standard-setting methods in this course. They conducted the standard-setting session during their final exam for partial credit.

Items

Items were nine four-option multiple-choice items taken from the final exam of these students. The items did not have "none of the above" or "all of the above" as response options. These nine items were also part of the final exam for previous students taking the same course. There were data on 274 past examinees (not including the 22 judges) responding to these nine items. Item difficulties estimated from these 274 past examinees were used as empirical p-values to evaluate intrajudge inconsistency for both the Angoff and Nedelsky methods.

Procedure

After turning in the final exam, the students were given the nine items from the exam on a separate sheet of paper with the correct answers marked. They were asked to apply the Angoff procedure on these items first. Instead of rating for minimum competency, they were instructed to estimate the probability an average student in this course would correctly answer each of the nine items. After turning in the Angoff ratings, the student judges were given another sheet of paper containing the nine

items with the correct responses marked. The students were instructed to cross out the false responses they thought an average student in this course would be able to eliminate. The student judges were only asked to decide whether an average student would be able to eliminate a false alternative; they were not asked to calculate the Nedelsky rating for an item. This instruction was intended to focus the judges' attention on the evaluation of alternative responses.

Results

An average squared deviation of item ratings from corresponding empirical p-values was computed for both the Nedelsky and Angoff methods:

$$\hat{\sigma}^2_i = \Sigma(X_i - Y_i)^2 \div n_i$$

where X_i is the Angoff or Nedelsky rating based on the 22 judges for item i; Y_i is p-value based on 274 past examinees for item i; n_i is number of items, which is nine.

This variance estimate which is like the Euclidean distance was used to measure intrajudge inconsistency. $\hat{\sigma}^2_i$ for the Angoff method was 0.06 and for the Nedelsky method was 0.02. Taking the square root of these variances yielded 0.24 and 0.14 which meant the average deviation of cutscores from actual p-values was 2.16 items or 24% of the nine items for the Angoff method and 1.08 or 12% of the items for the Nedelsky method. Thus, intrajudge inconsistency for the Angoff method was three times that of the Nedelsky method. To test the first hypothesis that Nedelsky ratings had higher internal consistency or lower intrajudge

inconsistency, these two variance estimates were compared by dividing the larger by the smaller of the two to yield an F-ratio. $F(8, 8) = 3$ was not significant, $p<.05$. The F-test was not significant partly because the sample size of nine items was extremely small.

The average of the p-values of the nine items based on 274 past examinees was 0.58. This number was almost identical to the Nedelsky cutscore of 0.57. The Angoff cutscore (.71) however, largely deviated from this empirical value. This evidence further supported the first hypothesis that the Nedelsky procedure resulted in smaller intrajudge inconsistency.

The second hypothesis that the Angoff cutscore was higher than the Nedelsky cutscore was tested and supported by a dependent t-test comparing the two cutscores provided by the same 22 judges. The mean rating of the 22 judges or cutscore using the Angoff method (0.71 or 71%) was significantly higher than that using the Nedelsky (0.57 or 57%); $t(20) = 5.44$, $p<.01$.

Additional t-tests were conducted to compare each of the nine pairs of item ratings by the same 22 judges. These results as well as means and standard deviations of the two methods and empirical p-values are reported in Table 2. Four of the nine rating comparisons were significant. They were from items having low p-values. These results lend partial support to the third hypothesis that there were larger differences between the Angoff and Nedelsky methods (Angoff having higher standards) for more difficult items. To further test this last hypothesis,

information reported in Table 2 was broken down for judges who answered the items correctly versus those who answered the items incorrectly. These results are contained in Table 3. Although for judges who correctly answered the items and those who did not, their Angoff ratings were higher than their Nedelsky ratings, the differences were much larger for judges who failed the items. These results supported the third hypothesis.

Conclusion

Intrajudge inconsistency was lower for the Nedelsky than the Angoff methods. The Nedelsky standard-setting procedure is more internally consistent possibly because judges use the same source of information in making decisions. In this method, judges are instructed to cross out false alternatives that they think a minimally competent examinee is capable of eliminating. Corresponding to this specific instruction, judges' cognitive process is focused on studying the plausibility and similarity among the alternatives and nothing else. In fact, in the reported standard-setting practice, judges did not have to divert their attention to actually calculating the Nedelsky rating for each item. All they did was to cross out the response options which they thought the targeting examinees were able to eliminate. Future standard-setting activities can adopt this method to focus judges' attention on the most important task. The low intrajudge inconsistency of the Nedelsky method might also be explained by the counter-balancing effect of multiple-decision making. For a four-option item, the MPL derived from

the Nedelsky procedure is the sum of three decisions. Intrajudge inconsistency associated with the final MPL may be ameliorated to the extent that not all three decisions are incorrect. For the same item, the Angoff MPL represents one decision with one error which can not be adjusted.

However, the multiple Nedelsky decisions associated with an item are dichotomous, limiting their counter-balancing potential. The dichotomous decisions also result in a fixed number of MPL values. For a four-choice item, there are four possible MPL's -- 0.25, 0.33, 0.5, and 1.0. These numbers or any fixed numbers do not represent the reality where probability ranges from 0.25 (for guessing) to 1 that a targeting examinee can correctly answer the item. The discreteness of the probability values produced by the Nedelsky method has been criticized (e.g., Brennan & Lockwood, 1981). Researchers have suggested modifications of the Nedelsky method that produces more continuous probability values (Gross, 1983; Saunders et al., 1981). One improvement would be to change the dichotomous decision regarding an examinee's ability to eliminate a distractor into a probability decision. The MPL for an item will be the sum of the probabilities of successfully eliminating all the distractors plus one (for guessing) divided by the number of response options. For a four-option item, the MPL will be the sum of the probabilities associated with the three distractors plus one divided by four. This modification of the Nedelsky method will produce a continuous MPL ranging from .25 to 1.0. More importantly, the counter-balancing power of the

multiple decisions will be maximized.

The standard derived from the Angoff procedure was higher than that derived by the same judges using the Nedelsky procedure. This finding is consistent with a seeming majority of the results in the literature. A more important finding, however, is that the Nedelsky method yields a cutscore closely approaching the actual performance of the examinees whereas the Angoff cutscore deviates greatly from examinees' actual performance level. At least for the short-answer items as those employed in this study, it is concluded that the Nedelsky method produces more adequate and more consistent standards than the Angoff method.

This study also demonstrates that the difference between the Nedelsky and Angoff methods changes as a function of judges' item-related knowledge. As hypothesized, the difference between the Nedelsky and Angoff standards increases for judges who have difficulties answering the items correctly themselves. This finding confirms the influence of judges' content knowledge in standard-setting. Logically, judges set higher standards for items of which they possess the underlying knowledge and set lower standards for items of which they lack the underlying knowledge. This influence is more pronounced in the Nedelsky procedure in which going through response alternatives subjects judges' item-related knowledge to a direct test. When judges do not know the answer to an item, an awareness of the lack of knowledge brought about or reinforced by the Nedelsky procedure

makes judges set a lower estimate of item difficulty. On the other hand, judges who possess the underlying knowledge of an item can come to view the item as more difficult as they go through the "tricky" alternatives. In the Angoff method which does not require a rigorous evaluation of the plausibility and similarity among alternative responses, judges' decisions may be predominantly driven by the plausibility of the correct answer that is presented to them. Without the scrutiny of the alternatives of an item, judges' potential lack of the underlying knowledge could escape their awareness. Some of the nearly plausible distractors of an item can indeed represent elements of the underlying knowledge of which judges are less confident. The absence of a close scrutiny of these alternatives and the presence of the reasonableness of the correct answer could make the judges overestimate the easiness of the item and consequently set a higher standard. Thus, when an item presents a potential challenge to the judges, the challenge has a higher chance of being factored into a Nedelsky than an Angoff decision. When judges have no difficulty with an item as shown, in this study, by their correct response to the item, the difference between the Angoff and Nedelsky decisions is reduced. However, even in this situation, the Nedelsky procedure still yields a lower standard than the Angoff method because similarity of alternatives adds challenge to the item and the Nedelsky method which is designed to tune in to such similarities is more likely to factor in this added difficulty. As Burton (1978) and Gross (1982) have pointed

out, the Nedelsky method correctly addresses the fact that multiple-choice item difficulty is a function not only of the complexity of the tested concept but also of the plausibility of the distractors.

The pronounced influence of judge competency on standard-setting in the Nedelsky method seems to be a desirable feature in the context of the present study. This conclusion is supported by the finding that the Nedelsky method in this study yielded a standard that was almost identical to the actual performance of the examinees. Additionally, this conclusion is supported by the finding that judges' item performance also closely matched the p-values of the targeting examinees. When judges' item difficulty represents that of the examinees, a cutscore influenced by judges' test performance shall be adequate for the performance of the examinees. However, this conclusion is limited to the context of this study where, in a way, judges were asked to set a standard for themselves. The ability level of an "average examinee" stands closer to the performance level of the judges than that of a minimally competent examinee. Thus, this conclusion is not directly generalizable. However, even though the absolute differences among item difficulties perceived by the judges are different than those experienced by minimally competent examinees, the relative differences among item difficulties can be expected to be the same for both judges and examinees. Thus, if judges can successfully determine the level of <u>minimum competency</u> in relation to <u>competency</u> which the judges

are selected to represent, estimating item difficulty for the former constitutes subtracting a constant from the item difficulty of the latter. The constant represents the distance between minimum competency and competency. In fact, when judges set standards, they consciously or unconsciously engage themselves in such a process of inferring the item difficulty from their experience or competency to that of the targeting examinees or minimum competency.

This study shows a clear superiority of the Nedelsky standard-setting method over the Angoff method -- the Nedelsky item ratings match the empirical p-values much more closely than the Angoff ratings. The superior result of the Nedelsky standard-setting method is attributed to the evaluation of the plausibility and similarity of the distractors. This activity forces judges to use a more consistent source of information in reaching decisions. It also probes judges' knowledge underlying an item so that judges gravitate toward factoring into their MPL decisions possible difficulties they themselves experience with the item. However, this study did not empirically test these possible contributors to the observed differences between the Nedelsky and Angoff methods. Like any non-experimental research, the present study suffers from the weakness of trying to infer, inversely, from the effect to its possible causes. Such inverse inference exists in almost all existing studies on standard-setting that tried to explain the differences between contrasting methods. Future research should aim at experimentally

manipulating the independent variables suggested by this study to determine their causal relations to standard-setting.

References

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Baron, J. b., Rindone, D. A., & Prowda, P. (April, 1981). Will the "real" proficiency standard please stand up? Paper presented at the annual meeting of the New England Educational Research Organization, Lenox, Massachusetts.

Behuniak, Jr., P., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. Educational and Psychological Measurement, 42(1), 247-255.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4(2), 219-240.

Burton, N. W. (1978). Societal standards. Journal of Educational Measurement, 15, 263-271.

Busch, J. C., & Jaeger, R. M. (1986, April). Judges' background, attitudes and information as concomitants of recommended test standards. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. Journal of

Educational Measurement, 27(2), 145-163.

Chang, L., Dzuiban, C., Hynes, M., Olson, A. (1994, April). Use generalizability theory to assess measurement errors of Angoff cutting scores and mean rater actual scores. Paper presented at the Annual Convention of the American Educational Research Association, New Orleans.

Cross, L. H., Frary, R. B., Kelly, P. P., Small, R. C., & Impara, J. C. (1985). Establishing minimum standards for essays: Blind versus informed reviews. Journal of Educational Measurement, 22, 137-146.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the national teacher examinations. Journal of Educational Measurement, 21(2), 113-129.

Gross, L. J. (1983, April). A refinement in the Nedelsky procedure for setting cutoff scores on credentialing examinations. Paper presented at the Annual Convention of the National Council on Measurement in Education, Montreal.

Gross, L. J. (1982). Standards and criteria: A response to Glass' criticism of the Nedelsky technique. Journal of Educational Measurement, 19(2), 159-161.

Halpin, Gerald., Sigmon, G, & Halpin, Glennelle (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 43(1), 185-197.

Harasym, P. H. (1981). A comparison of the Nedelsky and modified Angoff standard setting procedure on evaluation outcome. Educational and Psychological Measurement, 41(3), 725-735.

Jaeger, R. M. (1982). An interactive structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4(4), 461-475.

Jaeger, R. M. (1989). Certification of student competence. In R.L. Linn (Ed.), Educational Measurement (3rd ed., pp.485-514). New York: Macmillan.

Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64(3), 425-461.

Livingston, S. A. (1982, March). Assumptions of standard setting methods. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. Applied Measurement in Education, 2(2), 121-141.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Pavia, R. E. A., & Vu, N. V. (1979, April). Standards for acceptable level of performance in an objectives-based medical curriculum: A case study. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. Educational Measurement: Issues and Practice, 10(2), 15-16, 22, 25.

Plake, B. S., Impara, J. C., & Potenza, M. T. (1994). Content specificity of expert judgments in a standard-setting study. Journal of Educational Measurement, 31(4), 339-347.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981, April). An empirical investigation of the Angoff, Ebel and Nedelsky standard setting methods. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Roth, R. (1987). The differences between teachers and teacher educators when judging the NTE professional knowledge test to determine a cut-score. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Mobile, Al.

Saunders, J. C., Ryan, J. P., & Huynh, H. (1981). A comparison of two approaches to setting passing scores based on the Nedelsky procedure. Applied Psychological Measurement, 5(2), 209-217.

Smith, R. L, & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. Journal of Educational Measurement, 25(4), 259-274.

Table 1
Comparisons Beteen the Nedelsky and Angoff Standard-Setting Methods

Study	Test, Judge, Design	Cutscore		S.D.	
		Nedelsky	Angoff	Nedelsky	Angoff
Baron, Rindone, & Prowda, 1981.	36 multiple-choice language test items and 65 multiple-choice math items. For language test, 11 and 10 public school teachers used the Nedelsky and Angoff methods; for math test, nine teachers used each method. Nested design.	Language: 18.0 Math: 42.0	21.0 38.0		
Behuniak, Archambault, & Gable, 1982.	9th grade reading and math tests; 30 multiple-choice (4-choice) items each. 27 professional staff in reading and math; reading: two groups of 3 and 4 judges each used the Nedelsky and two groups of 3 and 4 judges each used the Angoff methods; math: two groups of 3 judges each used the Nedelsky and two groups of 3 and 4 judges each used the Angoff methods. Nested design.	Reading: 52.0 34.4 Math: 83.2 70.9	 49.6 63.8 76.5 65.5	 5.35 0.81 10.86 4.13	 3.03 3.19 2.59 3.91
Brennan & Lockwood, 1984.	126 multiple-choice (4-choice) items in a health related area. Five undefined judges. Crossed design.	70.09	83.56	9.03	4.70
Cross, Impara, Frary, & Jaeger, 1984[a].	National Teacher Examination in math and elementary education; 60 odd-numbered items form 120 math items and 75 odd-numbered items from 150 elementary items; multiple-choice (5-choice) items. Five faculty members per test per method. Nested design.	Math: 50.3% Elementary: 42.03%	 44.9% 56.68%	 15.34% 16.99%	 9.34% 19.35%

Halpin, Sigmon, & Halpin, 1983.	Missouri College English Test, 90 multiple-choice items. Five doctoral students in English, five high school English teachers, and five faculty members. Crossed design.	Students: 46.61	51.14
		Teachers: 34.49	62.54
		Faculty: 46.80	54.38
Harasym, 1981.	Endocrinology exam; 1979 class: 60 multiple-choice (4- to 5-choice) and 126 multiple true-false (4- to 5-choice) items; 1980 class: 51 and 152 items of the two types; 1981 class: 50 and 187 items of the two types; multiple-choice items were rated by the Nedelsky and multiple true-false items were rated by the Angoff methods. Eight unit managers served as judges. Nested design.	79 class: 56%	72%
		80 class: 58%	81%
		81 class: 59%	79%
Livingston & Zieky, 1989.	Basic Skills Assessment Tests in reading and mathematics, 4-option multiple-choice items, 65 and 70 items in reading and math. Reading test judges were teachers of English, reading, or language arts. Math test judges were math teachers. Three to five teachers from each of eight schools served as judges for each test. Four schools used the Nedelsky and four schools used the Angoff methods resulting in four comparisons: a: Judges for both methods were teachers of 6th, 7th, and 8th grade students. b: Judges for both methods were teachers of 7th and 8th grade students. c: Nedelsky judges were from 7th, 8th, and 9th grades and Angoff judges were from 7th and 8th grades. d: Nedelsky judges were 8th grade teachers and Angoff judges were 7th and 8th grade teachers. Nested design.	Reading: a: 38.1	25.9
		b: 36.5	32.9
		c: 46.0	43.2
		d: 46.7	37.6
		Math: a: 25.0	24.6
		b: 27.7	34.7
		c: 29.8	33.9
		d: 42.2	48.6

Poggio, Glasnapp, & Eros, 1981.	Kansas Competency-Based Test in reading and math for grades 2, 4, 6, 8, and 11; multiple-choice (4-choice) items; 45 items for grade 2, 60 items for grades 4, 6, and 8, and 57 items for grade 11. Teacher judges ranging from 24 to 41 per test per method. Nested design.	Reading				
		Grade 2:	19.1	36.4	2.7	4.5
		4:	25.2	42.4	3.3	11.2
		6:	24.2	43.3	3.7	7.8
		8:	25.0	42.3	2.4	9.8
		11:	23.3	41.5	2.1	7.8
		Math, 2:	18.0	39.1	2.9	3.8
		4:	25.1	45.6	3.6	8.7
		6:	25.5	42.9	2.6	7.5
		8:	25.0	38.2	2.9	10.1
		11:	20.5	36.9	2.9	10.3
Smith & Smith, 1988.	High school reading comprehension test, 64 multiple-choice (4-choice) items. 16 and 15 teachers or reading specialists used the Nedelsky and Angoff methods. Nested design.		71%	69%	8%	10%

Note. Studies are identified from PsycLit, 1974 to current, and ERIC, 1969 to current. Cutscores and standard deviations (S.D.) are presented in numbers of items (without %) or in percentages of items (with %).
*Results are from the first session before judges adjusted their ratings based on normative data. The Nedelsky ratings are uncorrected for guessing.

Table 2

Item Difficulty (P), Minimum Pass Level (MPL), Standard Deviation (SD), and T-Test Comparing Angoff and Nedelsky MPL

Item	P1	P2	Angoff		Nedelsky		T-Test
			MPL	SD	MPL	SD	
1	.76	.53	.73	.21	.67	.30	1.13
2	1.00	.76	.80	.15	.85	.26	1.07
3	.81	.61	.68	.21	.58	.26	1.64
4	.81	.62	.67	.22	.54	.28	2.60*
5	.48	.36	.65	.22	.44	.25	3.24*
6	.29	.11	.74	.17	.33	.10	7.35*
7	.95	.82	.73	.19	.57	.26	2.70*
8	.57	.66	.69	.22	.56	.33	1.90
9	.67	.76	.68	.25	.61	.29	1.06

Note. P1 = Item difficulty based on the 22 student judges who provided the Angoff and Nedelsky ratings. P2 = Item difficulty based on 274 past examinees.

*p<.05, two-tailed test.

Table 3

Number of Judges (n_j), Minimum Pass Level (MPL), and Standard Deviation (SD) from Judges Who Answered the Items Correctly and Judges Who Answered the Items Incorrectly

	Judges Who Answered the Items Correctly					Judges Who Answered the Items Incorrectly				
		Angoff		Nedelsky			Angoff		Nedelsky	
Item	n_j	MPL	SD	MPL	SD	n_j	MPL	SD	MPL	SD
1	16	.77	.20	.77	.27	5	.60	.20	.37	.13
2	21	.80	.15	.85	.26	0	--	--	--	--
3	17	.66	.21	.65	.23	4	.76	.19	.29	.05
4	17	.72	.21	.60	.28	4	.49	.16	.31	.04
5	10	.67	.18	.62	.27	11	.62	.26	.29	.08
6	6	.62	.21	.43	.11	15	.79	.12	.29	.07
7	20	.73	.19	.58	.26	1	.80	.00	.25	.00
8	12	.78	.22	.76	.31	9	.57	.15	.29	.08
9	14	.72	.27	.75	.26	7	.61	.21	.34	.11